Homework #2

See Canvas for HW #2 assignment due date. Total points without optional question: A1(a-b), A2(a-e), B1(a-b), B2(a-e), so 14 parts $*$ 2 = 28 total points. Optional question, if you choose to grade it: B3(1-5), so 5 parts $*$ 2 $=$ 10 points. So, if you choose to grade optional question, 28 $+$ 10 points $=$ 38 points.

A. Theoretical Problems

Problem A.1

Keep in mind that during class, we kept our equations in the 1,..,p-1 space for predictors. We will use this HW to get familiar with the other very common way of doing it.

Matrices and vectors will play an important role for us in linear regression. Let's review some matrix theory as it might relate to linear regression.

Consider the system of linear equations

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j} + \varepsilon_i,$$

for $i = 1, \ldots, n$, where n is the number of data points (measurements in the sample), and $j = 1, \ldots, p$, where

1. $p + 1$ is the number of parameters in the model.
2. Y_i is the i^{th} measurement of the *response variable*.
3. $X_{i,j}$ is the i^{th} measurement of the j^{th} *predictor variable*.
4. ε_i is the i^{th} *error term* and is a random variable (often assumed to be $N(0, \sigma^2)$).
5. β_j are *unknown parameters* of the model, ($j = 0, \ldots, p$). We hope to estimate these, which would help us characterize the relationship between the predictors and response.

(a) Write the equation above in matrix vector form. Call the matrix including the predictors X, the vector of Y_is $\mathbf{Y}$, the vector of parameters β, and the vector of error terms ε. (This is more LaTeX practice than anything else...)

(b) In class, we will find that the OLS estimator for β in MLR is $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{Y}$.

1. What condition must be true about the columns of X for the "Gram" matrix $X^T X$ to be invertible?
2. What does this condition mean in practical terms?
3. Suppose that the number of measurements (n) is less than the number of model parameters ($p + 1$). What does this say about the invertibility of $X^T X$? What does this mean on a practical level?
4. What is true about about $\widehat{\boldsymbol{\beta}}$ if $X^T X$ is not invertible?

Problem A.2

In class, we defined the *hat* or *projection* matrix as

$$H = X(X^T X)^{-1} X^T.$$

The goal of this question is to use the hat matrix to prove that the fitted values, $\widehat{\mathbf{Y}}$, and the residuals, $\widehat{\boldsymbol{\varepsilon}}$, are uncorrelated. We will do it in steps, and *some* of the proofs will only be required for STAT 5010 students. Note that STAT 4010 students are asked to answer part (e), as to why this result has practical importance.

(a) Show that $\widehat{Y} = HY$. That is, H "puts a hat on" Y.

(b) Show that H is symmetric: $H = H^T$.

(c) Show that $H(I_n - H) = 0_n$, where 0_n is the zero matrix of size $n \times n$.

(d) Stating that $\widehat{\mathbf{Y}}$ is uncorrelated with $\widehat{\boldsymbol{\varepsilon}}$ is equivalent to showing that these vectors are orthogonal.* That is, we need to show that their dot product is zero:

$$\widehat{\mathbf{Y}}^T \widehat{\boldsymbol{\varepsilon}} = 0.$$

Prove this result.

(e) Why is this result important in the practical use of linear regression?

B. Computational Problems

Problem B.1

Let $X_1, \dots, X_{30} \overset{iid}{\sim} N(1, 9)$. The formula for a 90% confidence interval for μ is

$$\bar{X} \pm 1.64 \frac{\sigma}{\sqrt{n}}.$$

Let's conduct a simulation to confirm the coverage of this confidence interval.

(a) Generate $m = 500$ random samples of size $n = 30$ from $N(1, 9)$ and calculate the 90% confidence interval for each. Don't print anything.

(b) Estimate the coverage by finding the number of intervals that cover the true mean, and dividing my m.

Problem B.2

(a) Load the "gala" dataset, and describe the variables.

(b) Use ggplot() to explore the relationship between the Species variable (response) and Endemics, Elevation, Nearest, and Adjacent (predictor variables). You might do so by creating four separate scatter plots. Do these relationships look linear? Does the variability in Species change as a function of any of the predictors? Are there any outliers in any of the plots?

(c) Perform a linear regression with Species as the response and Endemics, Elevation, Nearest, and Adjacent as predictors. Interpret the parameter estimate associated with Endemics (assume, for the moment, that the model is correct, and so the interpretation holds).

(d) Calculate the residual sum of squares, and the total sum of squares for this model. Then, use these calculations to verify the Multiple R-squared calculation in the summary from the previous part. Interpret R^2 (assume, for the moment, that the model is correct, and so the interpretation holds).

(e) Plot the residuals vs the fitted values. Based on what we've discussed in class (and a question from Section A of this homework!), what do you expect to see in this plot? Do you see what you expect to see? If not, what does that mean?

Problem B.3 ATTENTION!!!: THIS PROBLEM IS OPTIONAL. IF YOU DON'T DO IT, SIMPLY IGNORE IT IN YOUR SELF-GRADING TOTAL. IF YOU DO IT, YOU CAN CHOOSE WHETHER TO INCLUDE IT IN YOUR SELF-GRADING OR NOT.

Here's a procedure for calculating a two-sample bootstrap hypothesis test. You will apply this procedure on real data below.

Let $X_1, \dots, X_{n_1}$ be an iid sample from population #1, with unknown mean μ_1 and known standard deviation σ_1, and let $Y_1, \dots, Y_{n_2}$ be an iid sample from population #2, with unknown mean μ_2 and known standard deviation σ_2. Suppose we want to conduct a hypothesis test of the sort:

$$H_0 : \mu_1 - \mu_2 = 0 \;\; vs \;\; H_1 : \mu_1 - \mu_2 \geq 0.$$

The following algorithm has been suggested for a bootstrap test.

1. Calculate the test statistic

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

2. Let $\bar{z}$ be the mean of the combined data sets. Create two new data sets, $x'_1, \dots, x'_{n_1}$ and $y'_1, \dots, y'_{n_2}$ that are the original data sets centered at $\bar{z}$.
3. Draw B random bootstrap samples of size n_1 from $x'_1, \dots, x'_{n_1}$ and of size n_2 from $y'_1, \dots, y'_{n_2}$. The result will be two matrices, x^* and y^*; x^* will containin columns of bootstrap samples from sample #1, and y^* will contain columns of bootstrap samples from sample #2.
4. Then, for each bootstrap sample pair, calculate

$$t^* = \frac{\bar{x^*}_j - \bar{y^*}_j}{\sqrt{\sigma_1^{*^2}/n_1 + \sigma_2^{*^2}/n_2}},$$

where $\bar{x^*}_j$ is the sample mean of the j^{th} bootstrap sample from sample #1, and $\bar{y^*}_j$ is the sample mean of the j^{th} bootstrap sample from sample #2. $\sigma_1^{*^2}$ and $\sigma_2^{*^2}$ are the corresponding variance estimates of the j^{th} bootstrap sample. t^* will be a vector of length B and will approximate the distribution of the test statistic t.

5. Estimate the p-value using

$$\frac{\#\text{ of times } \{t^* \geq t\}}{B}.$$

Appling this procedure to real data...

A tennis club has two systems to measure the speed of a tennis ball. The local tennis pros suspect one system, `speed1`, is consistently recording faster speeds. To test her suspicions, she sets up both systems and records the speed of 12 serves (three serves from each side of the court). The values are stored in the data frame `tennis`, with variables `speed1` and `speed2`. The recorded speeds are in kilometers per hour.

Does the evidence support the tennis pro's suspicion? Use the above bootstrap hypothesis testing procedure and $\alpha = 0.1$.